TADA-Denovo: Analysis of *De Novo* Mutations using the TADA Model

In our paper, we described the TADA model that combines multiple types of rare variant data from family and case-control studies to infer the risk genes. TADA can also be used to analyze *de novo* data alone, and its most useful application would be to assess the significance of genes with multiple *de novo* mutations in different categories (e.g. some nonsense and some missense). As the sample size of the *de novo* studies continues to grow, such genes will be increasingly common, and a powerful statistical test is essential. While the basic TADA model can be adopted for the *de novo* data alone (just setting the counts in other types of data 0), it is more convenient to have a stand-alone model for *de novo* analysis. We call this model TADA-Denovo, and describe how it works in this short document.

We first describe a naive approach to analyze genes with multiple *de novo* mutations, and explain why it is not desirable. Generally, if the multiple *de novo* events belong to different categories, extra care must be taken. To see this, suppose we have a gene with 2 *de novo* nonsense mutations and 1 *de novo* missense mutation in a sample of 1,000 trios. The nonsense mutation rate of this gene is 1×10^{-6} and the missense mutation rate is 2×10^{-5}. If we use the Poisson test for the nonsense mutation, the p-value is 2×10^{-6}: we observe two events, while expecting $2 \times 1000 \times 1 \times 10^{-6} = 0.002$ event. At the gene level, using the Poisson test, however, leads to a p-value of 1.2×10^{-5}: we observe three events in total, while expecting $2 \times 1000 \times (1 \times 10^{-6} + 2 \times 10^{-5}) = 0.042$ event. This is clearly counter-intuitive, as the extra *de novo* missense mutation actually reduces the significance of this gene. The problem with this approach is that the evidence from different types of *de novo* events are not weighed properly. The test would assign the same significance to genes with the same total number of *de novo* events, regardless of how these events are distributed across different categories. As we know intuitively, *de novo* nonsense mutations would carry higher weights than *de novo* missense mutations.

Another simple approach to combine multiple types of *de novo* mutations is to compute p-values of each type separately, then combine the p-values using some kind of meta-analysis, most obviously, Fisher's method of combining p-values. This approach, however, is also seriously flawed, and not appropriate as a general means of assessing genes with multiple events. In the example above, instead of having one *de novo* missense mutation, suppose the gene has no missense event (it still has two *de novo* nonsense). The Poisson tests on *de novo* nonsense and missense mutations give $p = 2 \times 10^{-6}$ and 1, respectively, and combining them using Fisher's method leads to $p = 2.8 \times 10^{-5}$. This is again counter-tuitive: given that *de novo* events are rare even for true risk genes, having no *de novo* missense event should not create such a penalty for this gene. The problem with the meta-analysis method is that: the power of *de novo* studies is not taken into account, thus an insignificant p-value is interpreted, incorrectly, as negative evidence instead of the lack of power.

Below we describe our Bayesian analysis of *de novo* data. At the first step, all *de novo* events in the data are divided into J different categories. Several schemes are possible, e.g. (1) nonsense and missense, or (2) loss-of-function (LoF), possibly damaging missense and probably damaging missense. TADA-Denovo analyzes each type of events separately, then combine the evidence in a Bayesian fashion. For the j-th category of a gene being tested, suppose it has $x^{(j)}$ *de novo* mutations out of a sample of N trios, and the mutation rate of this genes in this category is $\mu^{(j)}$. Based on the *de novo* part of the TADA model (Figure 2 in the paper), the likelihood is:

$$x^{(j)}|\gamma_j \sim \text{Pois}(2N\mu^{(j)}\gamma_j) \tag{1}$$

where γ_j is the relative risk of the *de novo* mutations of the j-th category. We are testing two models: the null model M_0 that the gene is not a risk gene, and the alternative model M_1 that it is. Under M_0, $\gamma_j = 1$. Under M_1, we assume a prior distribution of γ_j (conjugate prior of Poisson distribution):

$$\gamma_j|M_1 \sim \text{Gamma}(\bar{\gamma}^{(j)}\beta^{(j)}, \beta^{(j)}) \tag{2}$$

where $\bar{\gamma}^{(j)}$ is the prior mean of the relative risk and $\beta^{(j)}$ controls the variance of the prior. This allows us to compute the marginal likelihood:

$$P(x^{(j)}|M_0) = \text{Pois}(x^{(j)}|2N\mu^{(j)}) \tag{3}$$

$$P(x^{(j)}|M_1) = \int P(x^{(j)}|\gamma_j)P(\gamma_j|M_1)d\gamma_j = \text{NegBin}\left(x^{(j)}|\bar{\gamma}^{(j)}\beta^{(j)}, \frac{2N\mu^{(j)}}{\beta^{(j)} + 2N\mu^{(j)}}\right) \tag{4}$$

The inference on the role of the gene is primarily based on the Bayes factor (BF), which is the product of the BFs computed from each category of events:

$$B = \prod_{j=1}^{J} \frac{P(x^{(j)}|M_1)}{P(x^{(j)}|M_0)} \tag{5}$$

TADA-Denovo also computes the p-value of the BF of a gene. To do this, it samples the number of *de novo* events in each category under the null model M_0 using Equation 1 with $\gamma_j = 1$. Then the BFs of all sampled genes (using all categories) are computed, which form the null distribution.

One issue of applying TADA-Denovo is to choose the values of the parameters of the prior distribution of relative risks, $\bar{\gamma}^{(j)}$ and $\beta^{(j)}$. We used a method of moment (MOM) estimation for these parameters in our analysis of ASD data, and the details can be found in Section 6 of Text S1 of the paper. Specifically, suppose we observe a total of $C^{(j)}$ *de novo* events in the j-th category across all genes in the human genome in a sample of N families, and a total of $M^{(j)}$ multiple-hit genes, i.e. genes sustaining more than one *de novo* events in the j-th category. The basic strategy is to set the values of the prior parameters so that the expected number of *de novo* events and that of multiple-hit genes match the observed values. The TADA software provides a function, `denovo.MOM()`, for the estimation. For the j-th type of events, it takes as input: the sample size N, the number of risk genes k (the same for all categories), the mutation rate of the j-th type of all genes $\mu^{(j)}$ (a vector), the total count of *de novo* events $C^{(j)}$, and the prior parameter $\beta^{(j)}$. The function computes two values: $\gamma^{(j)}$ that is consistent with the input, and $M_e^{(j)}$ the expected number of multiple-hit genes in the j-th category. Typically, one could choose a value of $\beta^{(j)}$ between 0.5 and 1, and then choose k so that $M_e^{(j)}$ is close to $M^{(j)}$ for all j's.

In our analysis of ASD data, we focus on two mutational categories: LoF and probably damaging missense (mis3) according to PolyPhen 2. The parameters estimated from the MOM approach are:

$$k = 1000 \qquad \beta^{\text{LoF}} = \beta^{\text{mis3}} = 1 \qquad \bar{\gamma}^{\text{LoF}} = 20 \qquad \bar{\gamma}^{\text{mis3}} = 4.7 \tag{6}$$

Using these parameters, we could obtain the p-values for the examples introduce earlier, specifically,

- Example 1: a gene with 2 *de novo* nonsense mutation and 1 *de novo* missense mutation, $p < 5 \times 10^{-8}$, which is much smaller than the p-value from the Poisson test using nonsense data alone.
- Example 2: a gene with 2 *de novo* nonsense mutation and 0 *de novo* missense mutation, $p = 2 \times 10^{-6}$, which is about the same as the p-value from Poisson test using nonsense data alone.

In either case, the p-value of the TADA-Denovo model is more reasonable than the naive Poisson test or meta-analysis discussed before.